REPORT DOCUMENTATION PAGE

AFRL-SR-BL-TR-98-

0207

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE 13 March 1997	3. REPORT TYPE AND DATES COVERED Final (01 Sep 93 – 31 Aug 96)

4. TITLE AND SUBTITLE Issues in Scaling Up Machine Learning	5. FUNDING NUMBERS F49620-93-1-0569
6. AUTHORS Dr. Michael J. Pazzani	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) University of California, Irvine Irvine, CA 92717	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) AFOSR/NM 110 Duncan Avenue, Room B-115 Bolling Air Force Base, DC 20332-8080	10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT Approved for Public Release	12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This grant has supported one graduate student researcher under the direction of Dr. Michael J. Pazzani investigating issues in improving the accuracy of machine learning systems. The classic machine learning paradigm for prediction has been to learn a set of decision structures or models from a training set" and select one for prediction on unseen "test data". Rather than select a single node from the set, the focus of this project's research has been to combine the prediction of the learned models to form an improved estimate. The two fronts of this research are regression and classification. In the realm of regression, the task is to predict a single continuous value for an example, .The majority of research in this area has focused on simple linear combination of the learned models. The nature of these weights may span from being highly regularized completely unconstrained. A set of weights is considered highly regularized if they are all positive, they sum to one, or they are uniform. Completely unconstrained weights have no restrictions) and may be derived by methods like ordinary least squares regression. The degree of regularization required depends on the particular regression problem. The project has developed a technique called PCRY, which automatically estimates the appropriate degrease regularization for a given data set. The basic idea is to use the eigen structure of the model predictions on the training data to derive a continuum of possible weight sets ranging front highly regularized to completely unconstrained. Cross-validation is used to estimate which weight set is most appropriate.

19980225 061

14. SUBJECT TERMS accuracy, regression, classification	15. NUMBER OF PAGES
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

F49620-
93-1-0569

Final Report: Issues in Scaling up Machine Learning

Michael J. Pazzani

March 13, 1997

This grant has supported one graduate student reseracher under the direction of Dr. Michael J. Pazzani invstigating issues in improving the accuracy of machine learning systems. The classic machine learning paradigm for prediction has been to learn a set of decision structures or models from a "training set" and select one for prediction on unseen "test data". Rather than select a single model from the set, the focus of this project's research has been to combine the predictions of the learned models to form an improved estimate. The two fronts of this research are regression and classification. What follows is a summary of the research for these two tasks.

In the realm of regression, the task is to predict a single continuous value for an example, $\mathbf{x}$. The majority of research in this area has focused on simple linear combinations of the learned models. For example, a combined estimate, $\hat{f}(\mathbf{x})$, can be derived by assigning weights, α, to each of the N learned models, $f_i(\mathbf{x})$, as follows:

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i f_i(\mathbf{x}).$$

The nature of these weights may span from being highly regularized to completely unconstrained. A set of weights is considered highly regularized if they are all positive, they sum to one, or they are uniform. Completely unconstrianed weights have no restrictions, and may be derived by methods like ordinary least squares regression. The degree of regularization required depends on the particular regression problem.

The project has developed a technique called PCR* [Merz and Pazzani, 1997, Merz and Pazzani, 1996] which automatically estimates the appropriate degree of regularization for a given data set. The basic idea is to use the eigen structure of the models' predictions on the training data to derive a continuum of possible weight sets ranging from highly regularized to completely unconstrained. Cross-validation is used to estimate which weight set is most appropriate. An

empirical evaluation of PCR* indicates that it indeed tends to choose the appropriate degree of regularization for a collection of data sets.

The current focus of the project's latests research is in the area of classification. Here, each learned model attempts to assign a class label to each example. Combining methods for classification typically assign weights to each learned model and take the "weighted majority", i.e.,

$$\hat{f}(\mathbf{x}) = arg \max_{c \in Y} \sum_{i=1}^{N} \alpha_i \|f_i(\mathbf{x}) = c\|$$

where Y is the set of possible classes, and $\|a = b\|$ is one of a is equal to b, and zero otherwise.

Initial research in this area [Merz, 1995] has indicated that the naive approach of taking the most frequent class (i.e., using uniform weights) is quite effective. However, when the learned models tend to make uncorrelated errors, more elaborate weighting methods may do better. In an attempt to explore weight sets analogous to those found in the regression task, two methods are currently under development.

The first is a direct extension of PCR* where each class is considered a separate (0/1) regression problem. The $i-th$ model's prediction for the $j-th$ class would be $f_{ij}(\mathbf{x})$. A zero for $f_{ij}(\mathbf{x})$ would indicate that example $\mathbf{x}$ is not predicted to be a member of class j, and a one would idicate the opposite. One weight set would be derived for each of the J classes as follows:

$$\hat{f}_j(\mathbf{x}) = \sum_{i=1}^{N} \alpha_{ij} f_{ij}(\mathbf{x}).$$

In deriving the combined estimate, $\hat{f}(\mathbf{x})$, a slightly more elaborate weighted majority scheme than given above is used. In this case, each learned model has a particular weight for each possible class. A preliminary evaluation of this extension of PCR* indicates that it also does a good job of choosing the appropriate weight set for a given data set.

The second approach being explored is based on a statistical procedure known as "correspondence analysis". Here the examples *and* the learned models are scaled into the same geometric space. In this representation, relationships between examples and learned models may be exploited so that a learned model's weight may be increased for examples on which it is likely to be accurate.

References

[Merz and Pazzani, 1996] Merz, C. and Pazzani, M. (1996). Handling redundancy in ensembles of learned models using principal components. In *Pro-*

ceedings of the National Conference on Artificial Intelligence, (workshop on Integrating Multiple Learned Models). IEEE Service Center.

[Merz, 1995] Merz, C. J. (1995). Dynamical selection of learning algorithms. In Fisher, D. and Lenz, H., editors, *Learning from Data: Artificial Intelligence and Statistics, 5.* Springer Verlag.

[Merz and Pazzani, 1997] Merz, C. J. and Pazzani, M. J. (1997). Combining neural network regression estimates with regularized linear weights. In Mozer, M., Jordan, M., and Petsche, T., editors, *Advances in Neural Information Processing Systems*, volume 9. The MIT Press.